Mnohonásobná a parciální korelace

Varianční, kovarianční a korelační matice

Nechť $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor. Označme

$\mu_i = E(X_i)$ střední hodnotu náhodné veličiny X_i,

$\sigma_i^2 = D(X_i)$ rozptyl náhodné veličiny X_i,

$\sigma_{ij} = C(X_i, X_j)$ kovarianci náhodných veličin X_i, X_j

$\rho_{ij} = R(X_i, X_j)$ koeficient korelace náhodných veličin X_i, X_j

Vektor $E(\mathbf{X}) = (\mu_1, \ldots, \mu_p)'$ se nazývá vektor středních hodnot náhodného vektoru **X**.

Čtvercová matice řádu p $var(\mathbf{X}) = (\sigma_{ij})_{i,j=1,\ldots,p}$ se nazývá varianční matice náhodného vektoru **X**.

Čtvercová matice řádu p $cor(\mathbf{X}) = (\rho_{ij})_{i,j=1,\ldots,p}$ se nazývá korelační matice náhodného vektoru **X**.

Je zřejmé, že varianční matice a korelační matice jsou symetrické.

Nechť $\mathbf{X} = (X_1, \ldots, X_p)'$ a $\mathbf{Y} = (Y_1, \ldots, Y_q)'$ jsou náhodné vektory.

Matice typu pxq $cov(\mathbf{X},\mathbf{Y}) = (C(X_i, Y_j))$ se nazývá kovarianční matice vektorů **X**, **Y**.

Matice typu pxq $cor(\mathbf{X},\mathbf{Y}) = (\rho(X_i, Y_j))$ se nazývá korelační matice vektorů **X**,**Y**.

Odhady vektoru středních hodnot, varianční a korelační matice jednoho náhodného vektoru X

Nechť $\mathbf{X}$ je náhodný vektor, který má p-rozměrné rozložení s vektorem středních hodnot $\boldsymbol{\mu}$, varianční maticí var($\mathbf{X}$) a korelační maticí cor($\mathbf{X}$). Nechť je dán náhodný výběr $\mathbf{X}_1 = (X_{11}, \ldots, X_{1p})', \ldots, \mathbf{X}_n = (X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení.

Nestranný odhad vektoru $\boldsymbol{\mu}$ je vektor výběrových průměrů $\mathbf{M} = (M_1, \ldots, M_p)'$, kde $M_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ je výběrový průměr j-tého výběru, j = 1, …, p.

Nestranný odhad matice var($\mathbf{X}$) je výběrová varianční matice $\mathbf{S} = (S_{ij}) = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{M})(\mathbf{X}_i - \mathbf{M})'$ řádu p.

Vychýlený odhad matice cor($\mathbf{X}$) je výběrová korelační matice $\mathbf{R} = (R_{ij})$, kde R_{ij} je výběrový korelační koeficient i-té a j-té složky vektoru $\mathbf{X}$, tedy

$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$, i, j = 1, …, p. (Je zřejmé, že diagonální prvky matice $\mathbf{R}$ jsou jedničky a matice $\mathbf{R}$ je symetrická.)

Příklad: U 28 náhodně vybraných osob byly zjišťovány tyto údaje:
Sex … 1 – muž, 2 – žena (mužů i žen bylo po 14)
výška (v cm), proměnná X_1
hmotnost (v kg), proměnná X_2
boty (číslo bot), proměnná X_3
Vypočtěte realizaci výběrové varianční matice a výběrové korelační matice. (Soubor udaje_o_lidech_1.sta)

Řešení:
Statistiky – Vícenásobná regrese - Proměnné Závislá X_3, nezávislé X_1, X_2 – OK – OK – Residua/předpoklady/předpovědi – Popisné statistiky – Další statistiky – Kovariance resp. Korelace.

Výběrová kovarianční matice

Proměnná	vyska	hmotnost	boty
vyska	112,8611	161,0926	41,45370
hmotnost	161,0926	248,4709	61,99206
boty	41,4537	61,9921	16,40608

Výběrová korelační matice

Proměnná	vyska	hmotnost	boty
vyska	1,000000	0,961979	0,963360
hmotnost	0,961979	1,000000	0,970948
boty	0,963360	0,970948	1,000000

Odhady kovarianční a korelační matice dvou náhodných vektorů X, Y

Nechť náhodný vektor **X** má p-rozměrné rozložení a nechť $\mathbf{X}_1, \ldots, \mathbf{X}_n$ je náhodný výběr z tohoto rozložení. Nechť náhodný vektor **Y** má q-rozměrné rozložení a nechť $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ je náhodný výběr z tohoto rozložení. Předpokládejme, že obě rozložení mají konečné druhé momenty. Nechť cov(**X**, **Y**) je kovarianční matice těchto vektorů a cor(**X**, **Y**) je korelační matice těchto vektorů. Označme $M_{Xj} = \frac{1}{n}\sum_{i=} \; _{ij}, j = 1,\ldots, p, M_{Yj} = \frac{1}{n}\sum_{i=} \; _{ij}, j = 1,\ldots, q$,

$\mathbf{M_X} = (M_{X1}, \ldots, M_{Xp})'$, $\mathbf{M_Y} = (M_{Y1}, \ldots, M_{Yq})'$.

Nestranným odhadem kovarianční matice cov(**X**, **Y**) vektorů **X**, **Y** je výběrová kovarianční matice vektorů **X**, **Y** definovaná vzorcem $\mathbf{S_{XY}} = (S_{ij}) = \frac{1}{n-1}\sum_{i=1}^{n} \mathbf{X}_i - \mathbf{M}_X \; \mathbf{Y}_i - \mathbf{M}_Y$, i = 1, …, p, j = 1, …, q.

Vychýleným odhadem korelační matice cor(**X**, **Y**) vektorů **X**, **Y** je výběrová korelační matice vektorů **X**, **Y** definovaná vzorcem $\mathbf{R_{XY}} = (R_{ij})$, kde R_{ij} je výběrový korelační koeficient i-té a j-té složky vektorů **X**, **Y,** i = 1, …, p, j = 1, …, q.

Příklad: Nechť vektor $\mathbf{X} = (X_1, X_2, X_3)'$ obsahuje údaje o výšce, hmotnosti a číslu bot mužů, vektor $\mathbf{Y} = (Y_1, Y_2)'$ obsahuje údaje výšce a hmotnosti žen. Vypočtěte realizace výběrové kovarianční a výběrové korelační matice vektorů **X**, **Y**. (Soubor udaje_o_lidech_2.sta)

Řešení:

Statistiky – Pokročilé lineární/nelineární modely – Obecné lineární modely – OK – Závislé proměnné: Vyska_z, Hmotnost_z – Spojité nezávislé proměnné: Vyska_m, Hmotnost_m, Boty_m – OK – na záložce Možnosti zaškrtneme Bez abs. členu – OK – na záložce Matice vybereme Kovariance resp. Korelace. Ve vzniklých tabulkách ponecháme pouze poslední dvě proměnné a první tři případy.

Výběrová kovarianční matice

Efekt	Sloup.4 Vyska_z	Sloup.5 Hmotnost_z
Vyska_m	10,81319	17,39560
Hmotnost_m	15,70879	15,22527
Boty_m	4,43407	5,13736

Výběrová korelační matice

Efekt	Sloup.4 Vyska_z	Sloup.5 Hmotnost_z
Vyska_m	0,467318	0,767160
Hmotnost_m	0,514047	0,508409
Boty_m	0,560289	0,662427

Koeficient mnohonásobné korelace a výběrový koeficient mnohonásobné korelace

Intenzitu lineární závislosti mezi náhodnou veličinou Y a náhodným vektorem $\mathbf{X} = (X_1, \ldots, X_p)'$ měříme pomocí koeficientu mnohonásobné korelace $\rho_{Y.\mathbf{X}}$. Jeho druhá mocnina je dána vzorcem

$$\rho_{Y.\mathbf{X}}^2 = \mathrm{cor}(Y, \mathbf{X})\, \mathrm{cor}(\mathbf{X})^{-1}\, \mathrm{cor}(\mathbf{X}, Y).$$

Má tyto vlastnosti:

a) $\rho_{Y.\mathbf{X}} \geq 0$

b) $\rho_{Y.\mathbf{X}} \geq |\rho(Y, X_i)|$ pro $\forall i = 1, \ldots, p$

c) $\rho_{Y.X_1 \ldots X_p} \geq \ldots \geq \rho_{Y.X_1X_2} \geq \rho(Y, X_1)$

d) $\rho_{Y.\mathbf{X}} = 1 \Leftrightarrow$ existují konstanty $\beta_0, \beta_1, \ldots, \beta_p$ tak, že $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

Nechť náhodný vektor $(Y, X_1, \ldots, X_p)'$ má $(p+1)$-rozměrné rozložení s koeficientem mnohonásobné korelace $\rho_{Y.\mathbf{X}}$. Nechť je dán náhodný výběr $(Y_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení. Pak jako odhad $\rho_{Y.\mathbf{X}}$ slouží výběrový koeficient mnohonásobné korelace $r_{Y.\mathbf{X}}$, jehož druhá mocnina je dána vzorcem

$$r_{Y.\mathbf{X}}^2 = \mathbf{R}_{Y\mathbf{X}}\, \mathbf{R}^{-1}\, \mathbf{R}_{\mathbf{X}Y},$$

kde $\mathbf{R}_{Y\mathbf{X}}$ je výběrová korelační matice veličiny Y a vektoru $\mathbf{X}$ (v tomto případě se redukuje na vektor $(r_{YX_1}, \ldots, r_{YX_p})$) a $\mathbf{R}$ je výběrová korelační matice vektoru $\mathbf{X}$.

Vlastnosti koeficientu mnohonásobné korelace se přenášejí i na výběrový koeficient mnohonásobné korelace.

Příklad: Při zkoumání závislosti hodinové výkonnosti dělníka (veličina Y – v kusech) na jeho věku (veličina X_1 – v letech) a době zapracovanosti (veličina X_2 – v letech) byly u 10 náhodně vybraných dělníků zjištěny tyto údaje:

Y	67	65	75	66	77	84	69	60	70	66
X_1	43	40	49	46	41	41	48	34	32	42
X_2	6	8	14	14	8	12	16	1	5	7

Vypočtěte výběrový koeficient mnohonásobné korelace r_{Y,X_1,X_2} popisující závislost hodinové výkonnosti dělníka na na jeho věku a době zapracovanosti.

Řešení:

Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2 – OK – OK.

Koeficient r_{Y,X_1,X_2} najdeme v záhlaví výstupní tabulky pod označením R = 0,54

Výsledky regrese se závislou proměnnou : Y (vykony delniku.sta)
R= ,54005243 R2= ,29165662 Upravené R2= ,08927280
F(2,7)=1,4411 p<,29913 Směrod. chyba odhadu : 6,6491

N=10	b*	Sm.chyba z b*	b	Sm.chyba z b	t(7)	p-hodn.
Abs.člen			86,74217	25,32397	3,425299	0,011056
X1	-0,550937	0,598452	-0,70031	0,76071	-0,920604	0,387883
X2	0,920415	0,598452	1,35062	0,87817	1,537994	0,167937

Jeho druhá mocnina (ozn. R2) nám říká, že variabilita výkonů dělníků je z 29% vysvětlena jejich věkem a dobou zapracovanosti.

Testování hypotézy o nezávislosti veličiny Y a vektoru X

Popis testu

Nechť náhodný výběr $(Y_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, X_{n1}, \ldots, X_{np})'$ pochází z (p+1)-rozměrného normálního rozložení, které má koeficient mnohonásobné korelace $\rho_{Y.\mathbf{X}}$. Musí platit $n > p+1$.

Testujeme hypotézu H_0: $\rho_{Y.\mathbf{X}} = 0$ proti H_1: $\rho_{Y.\mathbf{X}} \neq 0$. Vzhledem k tomu, že se jedná o výběr z (p+1)-rozměrného normálního rozložení, testujeme, zda existuje závislost mezi veličinou Y a vektorem **X**. (Je-li $\rho_{Y.\mathbf{X}} = 0$, pak z vlastnosti (b) plyne, že $\rho(Y,X_i) = 0$ pro všechna i = 1, …, p, tudíž náhodné veličiny Y a X_i jsou stochasticky nezávislé pro všechna i = 1, …, p.)

Testová statistika $F = \frac{n-p-1}{p} \cdot \frac{r_{Y.X}^2}{1-r_{Y.X}^2}$ se řídí rozložením F(p, n-p-1), pokud H_0 platí. Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty \rangle$.

Jestliže $F \in W$, H_0 zamítáme na hladině významnosti α.

Příklad

Předpokládáme, že údaje o výkonnosti 10 náhodně vybraných dělníků, jejich věku a době zapracovanosti představují číselné realizace náhodného výběru rozsahu 10 ze třírozměrného normálního rozložení. Na hladině významnosti 0,05 testujte hypotézu, že výkon dělníka nezávisí na jeho věku a době zapracovanosti.

Řešení:

Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2 – OK – OK.

Výsledky regrese se závislou proměnnou : Y (vykony delniku.sta)
R= ,54005243 R2= ,29165662 Upravené R2= ,08927280
F(2,7)=1,4411 p<,29913 Směrod. chyba odhadu : 6,6491

N=10	b*	Sm.chyba z b*	b	Sm.chyba z b	t(7)	p-hodn.
Abs.člen			86,74217	25,32397	3,425299	0,011056
X1	-0,550937	0,598452	-0,70031	0,76071	-0,920604	0,387883
X2	0,920415	0,598452	1,35062	0,87817	1,537994	0,167937

Hodnota testové statistiky pro test nevýznamnosti koeficientu mnohonásobné korelace $\rho_{Y,(X_1,X_2)}$ je 1,4411, počet stupňů volnosti čitatele je 2, jmenovatele 7, odpovídající p-hodnota je 0,2991, tedy na hladině významnosti 0,05 nezamítáme hypotézu, že výkon dělníka není závislý na jeho věku a době zapracovanosti.

Koeficient parciální korelace

Nechť Y, Z jsou náhodné veličiny a $\mathbf{X} = (X_1, \ldots, X_p)'$ je náhodný vektor. Korelační koeficient $\rho(Y,Z)$ udává míru těsnosti lineárního vztahu mezi veličinami Y a Z. Ta však může být ovlivněna i tím, že mezi veličinami $X_1, \ldots, X_p$ existují veličiny, které silně korelují jak s Y, tak se Z. Zajímá nás proto, jaká je „čistá“ korelace mezi Y a Z, když se eliminuje vliv náhodného vektoru **X**.

Pokud se omezíme na lineární vztahy, můžeme vliv vektoru **X** na veličinu Y popsat lineární regresní funkcí

$\hat{Y} = \alpha + \boldsymbol{\beta}'\mathbf{X}$, kde $\boldsymbol{\beta} = \mathrm{var}(\mathbf{X})^{-1}\,\mathrm{cov}(\mathbf{X},Y)$, $\alpha = E(Y) - \boldsymbol{\beta}'E(\mathbf{X})$.

Tu část veličiny Y, kterou vektor **X** nevysvětlí, si můžeme představit jako reziduum $Y - \hat{Y}$. Analogicky pro veličinu Z dostáváme

$\hat{Z} = \gamma + \boldsymbol{\delta}'\mathbf{X}$, kde $\boldsymbol{\delta} = \mathrm{var}(\mathbf{X})^{-1}\,\mathrm{cov}(\mathbf{X},Z)$, $\gamma = E(Z) - \boldsymbol{\delta}'E(\mathbf{X})$,

tudíž reziduum $Z - \hat{Z}$ chápeme jako tu část veličiny Z, kterou vektor **X** nevysvětlí.

Korelační koeficient mezi rezidui $Y - \hat{Y}$ a $Z - \hat{Z}$ se nazývá parciální korelační koeficient mezi náhodnými veličinami Y a Z při pevně daném vektoru **X** a značí se $\rho_{Y,Z.X}$. Tedy $\rho_{Y,Z.X} = \rho(Y - \hat{Y}, Z - \hat{Z})$. Počítá se podle vzorce

$$\rho_{Y,Z.X} = \frac{\rho(Y,Z) - \mathrm{cov}(Y,\mathbf{X})\,\mathrm{cor}(\mathbf{X})^{-}\,\mathrm{cov}(\mathbf{X},Z)}{\sqrt{\left(1 - \mathrm{cov}(Y,\mathbf{X})\,\mathrm{cor}(\mathbf{X})^{-}\,\mathrm{cov}(\mathbf{X},Y)\right)\left(1 - \mathrm{cov}(Z,\mathbf{X})\,\mathrm{cor}(\mathbf{X})^{-}\,\mathrm{cov}(\mathbf{X},Z)\right)}}.$$

Nechť náhodný vektor $(Y, Z, X_1, \ldots, X_p)'$ pochází z (p+2)-rozměrného rozložení, které má parciální korelační koeficient $\rho_{Y,Z.X}$. Nechť je dán náhodný výběr $(Y_1, Z_1, X_{11}, \ldots, X_{1p})', \ldots, (Y_n, Z_n, X_{n1}, \ldots, X_{np})'$ rozsahu n z tohoto rozložení. Musí platit $n > p+2$. Jako odhad $\rho_{Y,Z.X}$ slouží výběrový parciální korelační koeficient $r_{Y,Z.X}$:

$$r_{Y,Z.X} = \frac{r_{YZ} - \mathbf{S}_{YX}\mathbf{R}_{XX}^{-}\mathbf{S}_{XZ}}{\sqrt{\left(1 - \mathbf{S}_{YX}\mathbf{R}_{XX}^{-}\mathbf{S}_{XY}\right)\left(1 - \mathbf{S}_{ZX}\mathbf{R}_{XX}^{-}\mathbf{S}_{XZ}\right)}}$$

Testování hypotézy o nezávislosti veličin Y a Z při eliminaci vlivu vektoru X

Popis testu

Budeme předpokládat, že uvedený náhodný výběr pochází z (p+2)-rozměrného normálního rozložení.

Testujeme hypotézu H_0: $\rho_{y,z.\mathbf{x}} = 0$ proti H_1: $\rho_{y,z.\mathbf{x}} \neq 0$.

Vzhledem k tomu, že se jedná o výběr z normálního rozložení, testujeme, zda existuje závislost mezi Y a Z při eliminaci vlivu **X**.

Testová statistika $T_0 = \frac{r_{Y,Z.X}\sqrt{n-p-2}}{\sqrt{1-r_{Y,Z.X}^2}}$ se řídí rozložením t(n-p-2), pokud H_0 platí.

Kritický obor: $W = (-\infty, t_{1-\alpha/2}(n-p-2)) \cup (t_{1-\alpha/2}(n-p-2), \infty)$.

Jestliže $T_0 \in W$, H_0 zamítáme na hladině významnosti α

Příklad

Pro data z příkladu o výkonnosti dělníků vypočtěte výběrové parciální korelační koeficienty $r_{Y,X_1.X_2}, r_{Y,X_2.X_1}$, interpretujte je, porovnejte je s obyčejnými výběrovými korelačními koeficienty r_{YX_1}, r_{YX_2} a pro $\alpha = 0{,}05$ otestujte významnost uvedených parciálních korelačních koeficientů.

Výpočet pomocí systému STATISTICA

Nejprve vypočteme koeficient korelace mezi výkonem a věkem.
Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 2 seznamy – 1. seznam Y, 2. seznam X_1, X_2 – Výpočet.

Proměnná	X1
Y	0,2287

Dále vypočteme parciální korelační koeficient mezi výkonem a věkem při vyloučení vlivu doby zapracovanosti a otestujeme jeho významnost.
Statistiky – Základní statistiky/tabulky – Korelační matice – OK – na záložce Možnosti zaškrtneme Zobrazit r, úrovně p, počty N, na záložce Detaily zvolíme Parciální korelace – 1. seznam proměnných Y, X1, druhý seznam proměnných X2 – OK

Proměnná	Y	X1
Y	1,0000	-,3286
	p= ---	p=,388
X1	-,3286	1,0000
	p=,388	p= ---

Korelační koeficient mezi výkonem a věkem vyšel 0,2287, tedy s rostoucím věkem roste výkon. Parciální korelační koeficient mezi výkonem a věkem při vyloučení vlivu doby zapracovanosti vyšel -0,3286, tedy u dělníků se stejnou dobou zapracovanosti klesá s rostoucím věkem výkon.
Odpovídající p-hodnota je 0,388, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nevýznamnosti $\rho_{Y,X_1.X_2}$.

Nyní vypočteme koeficient korelace mezi výkonem a dobou zapracovanosti:

Proměnná	X2
Y	0,4538

Dále vypočteme parciální korelační koeficient mezi výkonem a dobou zapracovanosti při vyloučení vlivu věku pracovníka a otestujeme jeho významnost.

Proměnná	Y	X2
Y	1,0000	,5026
	p= ---	p=,168
X2	,5026	1,0000
	p=,168	p= ---

Korelační koeficient mezi výkonem a dobou zapracovanosti vyšel 0,4538, tedy čím delší doba zapracovanosti, tím lepší výkon dělník podává. Parciální korelační koeficient mezi výkonem a dobou zapracovanosti při vyloučení vlivu věku vyšel 0,5026, tedy u stejně starých dělníků je poněkud silnější přímá lineární vazba mezi výkonem a dobou zapracovanosti.

Odpovídající p-hodnota je 0,168, tedy na hladině významnosti 0,05 nezamítáme hypotézu o nevýznamnosti $\rho_{Y,X_2.X_1}$.

Mnohonásobná lineární regrese

Popis modelu mnohonásobné lineární regrese

Budeme zkoumat lineární závislost veličiny Y na p nezávisle proměnných veličinách $x_1, ..., x_p$. Omezíme se pouze na model tvaru

$$Y_i = \beta_0 + \beta_1 x_{il} + ... + \beta_p x_{ip} + \varepsilon_i,\ i = 1, ..., n.$$

Parametr β_0 interpretujeme jako teoretickou hodnotu závisle proměnné veličiny při nulových hodnotách všech nezávisle proměnných veličin. Parametr β_j, $j = 1, ..., p$ interpretujeme jako přírůstek teoretické hodnoty závisle proměnné veličiny odpovídající jednotkové změně j-té nezávisle proměnné veličiny při konstantní úrovni ostatních nezávisle proměnných.
Geometricky tento model představuje regresní nadrovinu. Lze ho formálně ztotožnit s lineárním regresním modelem z kapitoly „Jednoduchá lineární regrese“, kde položíme $f_1(x_i) = x_{i1}, ..., f_p(x_i) = x_{ip}$, $i = 1,..., n$. Dostáváme tedy maticový tvar $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$, kde regresní matice

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix},$$ přičemž $h(\mathbf{X}) = p+1 < n$ a $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I})$.

Všechny výsledky uvedené v přednáškách „Regresní analýza I“ a „Regresní analýza II“ zůstávají v platnosti.

Příklad:

Pro data z příkladu o výkonnosti dělníků sestavte regresní matici a vektor regresníc koeficientů.

Y	67	65	75	66	77	84	69	60	70	66
X_1	43	40	49	46	41	41	48	34	32	42
X_2	6	8	14	14	8	12	16	1	5	7

Řešení:

$$\mathbf{X} = \begin{pmatrix} 1 & 43 & 6 \\ 1 & 40 & 8 \\ 1 & 49 & 14 \\ 1 & 46 & 14 \\ 1 & 41 & 8 \\ 1 & 41 & 12 \\ 1 & 48 & 16 \\ 1 & 34 & 1 \\ 1 & 32 & 5 \\ 1 & 42 & 7 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Míra lineární závislosti veličiny Y na veličinách $x_1, ..., x_p$

Jak bylo uvedeno v předešlém textu, mírou těsnosti lineární závislosti náhodné veličiny Y na vektoru $\mathbf{X} = (X_1,..., X_p)'$ je koeficient mnohonásobné korelace $\rho_{Y.\mathbf{X}}$:

$\rho_{Y.\mathbf{X}}^2 = \mathrm{cor}(Y, \mathbf{X})\, \mathrm{cor}(\mathbf{X})^{-1}\, \mathrm{cor}(\mathbf{X}, Y)$, kde

$\mathrm{cor}(Y, \mathbf{X})$ je korelační matice veličiny Y a vektoru $\mathbf{X}$ (v tomto případě se redukuje na vektor $(\rho_{YX_1},..., \rho_{YX_p})$),

$\mathrm{cor}(\mathbf{X})$ je korelační matice vektoru $\mathbf{X}$.

Výběrovým protějškem koeficientu $\rho_{Y.\mathbf{X}}$ je výběrový koeficient mnohonásobné korelace $r_{Y.\mathbf{X}}$:

$r_{Y.\mathbf{X}}^2 = R_{Y\mathbf{X}}\, \mathbf{R}^{-1}\, R_{\mathbf{X}Y}$, kde

$R_{Y\mathbf{X}}$ je výběrová korelační matice veličiny Y a vektoru $\mathbf{X}$ (v tomto případě se redukuje na vektor $(r_{YX_1},..., r_{YX_p})$),

$\mathbf{R}$ je výběrová korelační matice vektoru $\mathbf{X}$.

V regresním modelu se mu říká index korelace. (V případě regresní přímky se jedná o obyčejný párový koeficient korelace r_{YX}.) Jeho kvadrát odpovídá indexu determinace v regresním modelu $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. Formálně je tedy celkový F-test rovnocenný s testem o nulové hodnotě koeficientu mnohonásobné korelace.

Stojí za zmínku, že vypočtená hodnota testové statistiky F by měla být aspoň 4x větší než příslušný kvantil Fisherova - Snedecorova rozložení, aby bylo možné prohlásit zvolený regresní model za skutečně kvalitní.

Posouzení vlivu jednotlivých nezávisle proměnných v modelu

Chceme-li porovnávat vliv, jaký mají proměnné $x_1, ..., x_p$ v modelu $\mathbf{Y = X\beta + \varepsilon}$, můžeme spočítat tzv. standardizované regresní parametry, kterým se také říká B-koeficienty. Zavedeme proto standardizované veličiny $z_i = \frac{Y_i - \bar{Y}}{s_Y}, v_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}}$,

$j = 1, ..., p, i = 1, ..., n$

a vytvoříme regresní model s těmito standardizovanými proměnnými. Odhady regresních parametrů v tomto novém modelu jsou B-koeficienty, které pak vyjadřují intenzitu vlivu jednotlivých nezávisle proměnných veličin na veličinu Y.

Příklad:

Pro data z příkladu o výkonnosti dělníků posuďte vliv věku a doby zapracovanosti na výkon dělníka pomocí standardizovaných regresních parametrů.

Řešení:

Statistiky – Vícenásobná regrese – Proměnné – Závislá proměnná Y, seznam nezáv. proměnných X1, X2 – OK – OK.

Výsledky regrese se závislou proměnnou : Y (vykony delniku.sta)
R= ,54005243 R2= ,29165662 Upravené R2= ,08927280
F(2,7)=1,4411 p<,29913 Směrod. chyba odhadu : 6,6491

N=10	b*	Sm.chyba z b*	b	Sm.chyba z b	t(7)	p-hodn.
Abs.člen			86,74217	25,32397	3,425299	0,011056
X1	-0,550937	0,598452	-0,70031	0,76071	-0,920604	0,387883
X2	0,920415	0,598452	1,35062	0,87817	1,537994	0,167937

Standardizované regresní parametry jsou uvedeny ve sloupci b^*. Pro věk má tento parametr hodnotu -0,5509 a pro dobu zapracovanosti 0,9204. V absolutní hodnotě je vyšší parametr pro dobu zapracovanosti, tedy tato proměnná má vyšší vliv na výkon než věk.

Použití parciálních korelačních koeficientů v modelu mnohonásobné lineární regrese

Uvažme model $Y_i = \beta_0 + \beta_1 x_{il} + ... + \beta_p x_{ip} + \varepsilon_i$, i = 1, ..., n. Druhá mocnina výběrového parciálního korelačního koeficientu $r_{Y,x_j.(x_1 \dots x_{j-1})}$, j = 2, ..., p se nazývá parciální index determinace. Lze ho interpretovat jako „čistý" přínos proměnné x_j do modelu, který obsahoval proměnné x_1, ..., x_{j-1}. Čím větší je závislost mezi x_j a $(x_1, ..., x_{j-1})$‘, tím menší se tento "čistý" přínos ukáže.

Výběrový parciální korelační koeficient $r_{Y,x_j.(x_1 \dots x_{j-1})}$ měří „čistou“ korelaci mezi Y a X_j, když se eliminuje vliv náhodného vektoru $(X_1, \dots, X_{j-1})$.

Protože v klasickém modelu lineární regrese je $S_T = S_R + S_E$, je pokles reziduálního součtu čtverců při zařazení nové proměnné do modelu roven růstu regresního součtu čtverců a naopak. Vzhledem k dříve zařazeným proměnným je tedy parciální index determinace mírou relativního zvýšení regresního součtu čtverců (poklesu reziduálního součtu čtverců) v důsledku zařazení nové proměnné.

Multikolinearita v modelu mnohonásobné regrese

O multikolinearitě hovoříme tehdy, když mezi některými sloupci regresní matice existuje silná lineární závislost, což svědčí o tom, že regresní model obsahuje nadbytečné vysvětlující proměnné.

Důsledky multikolinearity: matice **X'X** je blízká singulární matici => kvalita odhadu **b** je nízká => rozptyly odhadů b_0, b_1, ..., b_p jsou velké => intervaly spolehlivosti pro β_0, β_1, ..., β_p jsou široké.

Signály upozorňující na existenci multikolinearity:

- vysoké absolutní hodnoty výběrových korelačních koeficientů nezávisle proměnných (orientačně > 0,75)
- celkový F-test je významný, ale dílčí t-testy nikoliv.

Při použití statistického software lze informace o multikolinearitě získat pomocí koeficientu VIF (Variance inflation factor). Má-li tento koeficient hodnotu 1, pak příslušná nezávisle proměnná není korelovaná s ostatními nezávisle proměnnými, jestliže 1<VIF<5, pak existuje mírná korelace, pro VIF>5 vysoká korelace a pro VIF>10 extrémní multikolinearita.
V systému STATISTICA obdržíme VIF v Obecných regresních modelech. (Statistiky - Pokročilé lineární/nelineární modely – Obecné regresní modely). Po zadání závislé proměnné a nezávislých proměnných zvolíme Matice – Parciální korelace:

Příklad:

Pro data z příkladu o výkonnosti dělníků posuďte pomocí koeficientu VIF, zda proměnné věk a doba zapracovanosti mohou způsobit multikolinearitu v modelu $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Řešení:

Statistiky - Pokročilé lineární/nelineární modely – Obecné regresní modely – OK – Proměnné – Závislá Y, Spojité nezávisle proměnné X_1, X_2 – OK – Matice – Parcilání korelace.

Efekt	Toler.	Rozptyl Infl fak	R^2	Y Beta v	Y Parciál.	Y Semipar.	Y t	Y p
"X1"	0,282545	3,539258	0,717455	-0,550937	-0,328630	-0,292850	-0,920604	0,387883
"X2"	0,282545	3,539258	0,717455	0,920415	0,502564	0,489246	1,537994	0,167937

Koeficient VIF je 3,54, tedy mezi věkem a dobou zapracovanosti existuje jen mírná korelace.

Odstranění multikolinearity: do modelu se zařadí jen ty proměnné, které významně zlepšují odhad regresních parametrů.

Jednou z metod výběru nejlepší podmnožiny proměnných je step-wise regression (postupná regrese). Úkolem postupné regrese je najít ty prediktory, které co nejlépe vystihují variabilitu závisle proměnné veličiny a získat odhady parametrů lineární regresní funkce, s jejíž pomocí pak lze uspokojivě predikovat hodnoty závisle proměnné veličiny.

Postupná regrese se používá ve dvou variantách – dopředná (forward) a zpětná (backward).

Při metodě forward se prediktory postupně přidávají, při metodě backward se nejdříve zařadí všechny prediktory a pak se postupně odebírají.

Princip postupné regrese spočívá v tom, že regresní model je budován krok po kroku tak, že v každém kroku zkoumáme všechny prediktory a zjišťujeme, který z nich nejlépe vystihuje variabilitu závisle proměnné veličiny.

Zařazování prediktoru do modelu či jeho vylučování se děje pomocí sekvenčních F-testů.

Sekvenční F-test je založen na statistice F, která je podílem přírůstku regresního součtu čtverců při zařazení daného prediktoru do modelu a reziduálního součtu čtverců. Jestliže je tato statistika větší než hodnota zvaná „F to enter“ (česky „F na zahrnutí“, ve STATISTICE implicitně 1), je prediktor zařazen. Je-li statistika F menší než hodnota zvaná „F to remove“ (česky „F na vyjmutí“, ve STATISTICE implicitně 0), je již dříve zařazený prediktor z modelu vyloučen. Po vybrání proměnných do modelu jsou odhadnuty parametry lineární regresní funkce a kvalita regrese je posouzena indexem determinace.

Do modelu se postupně přidávají další proměnné, pokud se zvyšuje podíl vysvětlené variability hodnot veličiny Y.

Algoritmus postupné regrese:

1. krok: Vypočteme výběrové korelační koeficienty mezi závisle proměnnou Y a regresory $x_1, ..., x_p$. Do modelu vybereme ten regresor x_i, pro který je absolutní hodnota korelačního koeficientu největší.

2. krok: Sestavíme model $Y = \beta_0 + \beta_1 x_i$, MNČ odhadneme regresní koeficienty, vypočteme regresní a reziduální součty čtverců S_R a S_E a testové kritérium $F = \frac{S_R}{\frac{S_E}{n-2}}$. Pokud $F \geq F_{1-\alpha}(1,n-2)$, pak regresor x_i zařadíme do modelu.

3. krok: Vypočteme výběrové parciální korelační koeficienty mezi závisle proměnnou a regresory dosud nezařazenými do modelu a vyloučením vlivu regresoru x_i. Vybereme ten regresor x_j, pro který je absolutní hodnota parciálního korelačního koeficientu největší.

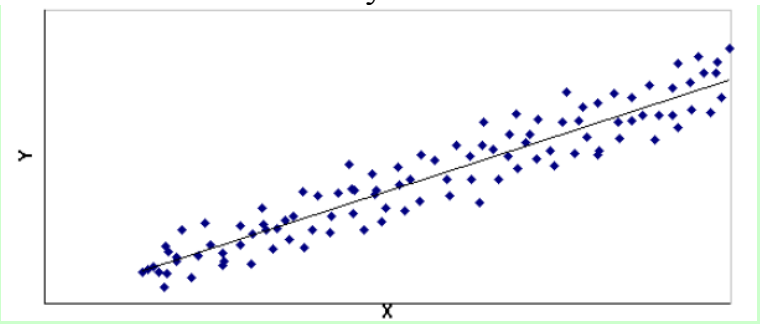
4. krok: Sestavíme model $Y = \beta_0 + \beta_1 x_i + \beta_2 x_j$, MNČ odhadneme regresní koeficienty, vypočteme regresní a reziduální součty čtverců S_R a S_E a testové kritérium $F = \frac{\Delta_R}{\frac{S_E}{n-3}}$, kde ΔS_R je přírůstek regresního součtu čtverců při zařazení regresoru x_j do modelu. Pokud $F \geq F_{1-\alpha}(1,n-3)$, pak regresor x_j zařadíme do modelu.

5. krok: Vypočteme výběrové parciální korelační koeficienty mezi závisle proměnnou a regresory dosud nezařazenými do modelu s vyloučením vlivu regresorů x_i a x_j a podle kroků 3 a 4 postupujeme dále, až vyčerpáme všechny regresory.

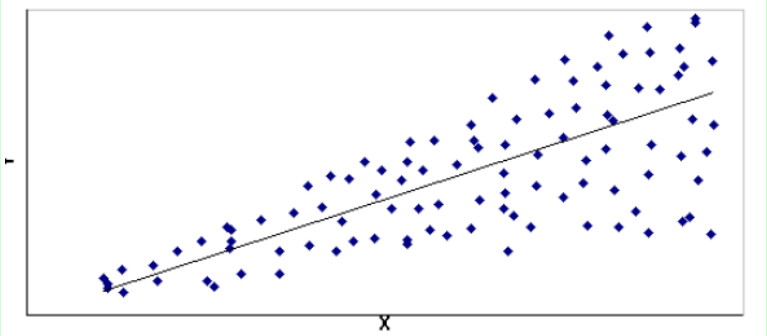
Postup při budování modelu mnohonásobné lineární regrese

1. Sestrojíme dvourozměrné tečkové diagramy dvojic (Y,x_j), $j = 1, ..., p$. Lze-li diagramem uspokojivě proložit přímku, svědčí to o tom, že Y lineárně závisí na x_j. Objeví-li se náhodný mrak bodů, Y na x_j záviset nebude. Obrazce jiných tvarů svědčí o problémech. Například trojúhelníkový tvar dvourozměrného tečkového diagramu indikuje heteroskedasticitu (tzn. že je porušena podmínka (d) v modelu klasické lineární regrese, tedy náhodné odchylky nemají týž rozptyl). Poučení o heteroskedasticitě lze nalézt např. v knize J. Hebák, J. Hustopecký: Vícerozměrné statistické metody s aplikacemi, SNTL 1987, Praha, kde je popsána zobecněná metoda nejmenších čtverců.

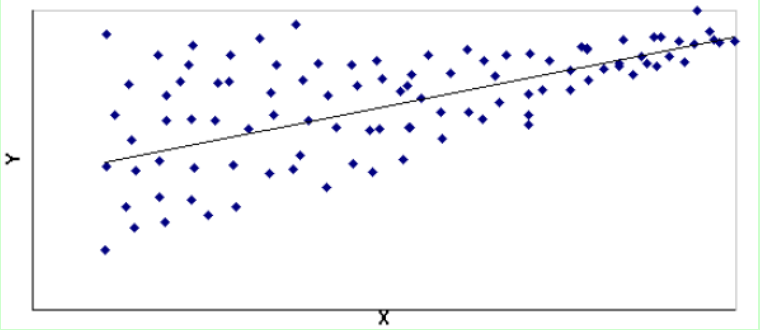
Ukázka homoskedastických dat:



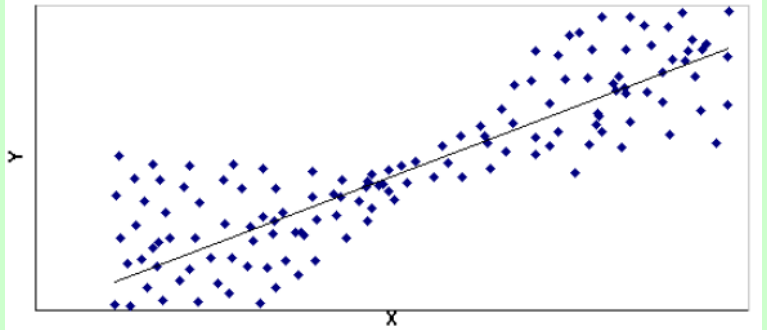
Ukázka dat s rostoucí heteroskedasticitou:



Ukázka dat s klesající heteroskedasticitou:



Ukázka dat s proměnlivou heteroskedasticitou:



2. Vypočteme výběrové párové korelační koeficienty, abychom posoudili sílu případné lineární závislosti Y na x_i . Dále vypočteme všechny výběrové parciální korelační koeficienty, abychom posoudili sílu „čisté" lineární závislosti mezi Y a x_j při vyloučení vlivu ostatních proměnných. Budou-li velké rozdíly mezi párovými a parciálními korelačními koeficienty, svědčí to o existenci multikolinearity.

3. V modelu $Y_i = \beta_0 + \beta_1 x_{il} + ... + \beta_p x_{ip} + \varepsilon_i$, i = l, ..., n získáme bodové a intervalové odhady regresních parametrů β_0, β_l, ..., β_p, index determinace, odhad rozptylu. Provedeme dílčí t-testy a celkový F-test. Vliv jednotlivých proměnných posoudíme pomocí B-koeficientů.

4. Z modelu vyloučíme ty nezávisle proměnné, pro něž byly dílčí t-testy nevýznamné.

Příklad

Šest studentů gymnázia absolvovalo čtyři testy, které měří následující veličiny: X_1 - přírodovědné vědomosti, X_2 – literární vědomosti, X_3 – schopnost koncentrace, X_4 – logické myšlení. Testy se hodnotí na škále od 1 do 10 (1 = špatný výsledek, 10 = výborný výsledek).

student	X_1	X_2	X_3	X_4
1	7	9	10	8
2	9	8	8	10
3	4	3	1	2
4	2	3	2	2
5	3	1	2	4
6	1	1	1	4

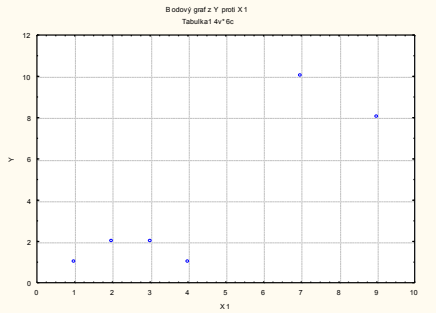
Zajímá nás, kolik bodů můžeme očekávat v testu koncentračních schopností studenta, jestliže známe výsledky testů pro literární schopnosti, přírodovědné schopnosti a logické myšlení.

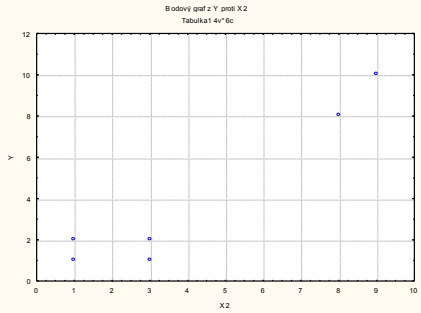
Řešení pomocí systému STATISTICA:

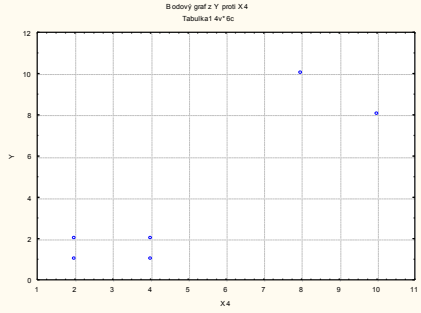
V tomto problému je proměnná X_3 závislá (označíme ji Y) a ostatní proměnné jsou nezávislé.

Sestavíme regresní model $Y_i = \beta_0 + \beta_1 x_{il} + \beta_2 x_{i2} + \beta_4 x_{i4} + \varepsilon_i$, i = 1, ..., 6.

Nejprve sestrojíme dvourozměrné tečkové diagramy vyjadřující závislost Y na X_1, X_2 a X_4.







Dále spočteme výběrové korelační koeficienty r_{Y,X_1}, r_{Y,X_2}, r_{Y,X_4} a výběrové parciální korelační koeficienty $r_{Y,X_1.X_2}$, $r_{Y,X_1.X_4}$, $r_{Y,X_2.X_1}$, $r_{Y,X_2.X_4}$, $r_{Y,X_4.X_1}$, $r_{Y,X_4.X_2}$.

Korelace (ctyri testy.sta)
Označ. korelace jsou v yznamné na hlad. p < ,05000
N=6 (Celé případy vynechány u ChD)

Proměnná	X1	X2	X4
Y	0,87	0,96	0,89

Vidíme, že korelace dvojic (Y, X_1), (Y, X_2), (Y, X_4) jsou vysoké.

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X1	Y
X1	1,0000	0,0273
Y	0,0273	1,0000

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X1	Y
X1	1,0000	0,4275
Y	0,4275	1,0000

Parciální korelace dvojice (Y, X_1) při vyloučení vlivu veličiny X_2 je pouze 0,0273 a při vyloučení vlivu veličiny X_4 je 0,4275, tedy mnohem slabší než párová korelace, která činila 0,87.

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X2	Y
X2	1,0000	0,8108
Y	0,8108	1,0000

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	X2	Y
X2	1,0000	0,8773
Y	0,8773	1,0000

Parciální korelace dvojice (Y, X_2) při vyloučení vlivu veličiny X_1 resp. X_4 je stále silná, jen o něco menší než párová korelace (ta byla 0,96).

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	Y	X4
Y	1,0000	0,5586
X4	0,5586	1,0000

	Parciální korelace (ctyri testy.sta) Označ. korelace jsou významné na hlad. p < ,05000 N=6 (Celé případy vynechány u ChD)	
Proměnná	Y	X4
Y	1,0000	0,6590
X4	0,6590	1,0000

Parciální korelace dvojice (Y, X_4) při vyloučení vlivu veličiny X_1 resp. X_2 je o dost menší než párová korelace (ta byla 0,89), ale pokles není tak výrazný jako u dvojice (Y, X_1) při vyloučení vlivu veličiny X_2 resp. X_4.

Z těchto analýz vyplývá, že největší roli v modelu lineární regresní závislosti Y na X_1, X_2 a X_4 bude hrát proměnná X_2, podstatně menší X_4 a role X_1 bude zřejmě jen nepatrná.

Metodou nejmenších čtverců získáme odhady regresních parametrů.

N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,98240301 R2= ,96511567 Upravené R2= ,91278918 F(3,2)=18,444 p<,05187 Směrod. chyba odhadu : 1,1664					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(2)	Úroveň p
Abs.člen			-1,08961	0,941927	-1,15679	0,366858
X1	-0,299065	0,368366	-0,38391	0,472872	-0,81187	0,502130
X2	0,864242	0,316998	0,97862	0,358949	2,72633	0,112320
X4	0,445257	0,271142	0,53513	0,325873	1,64215	0,242263

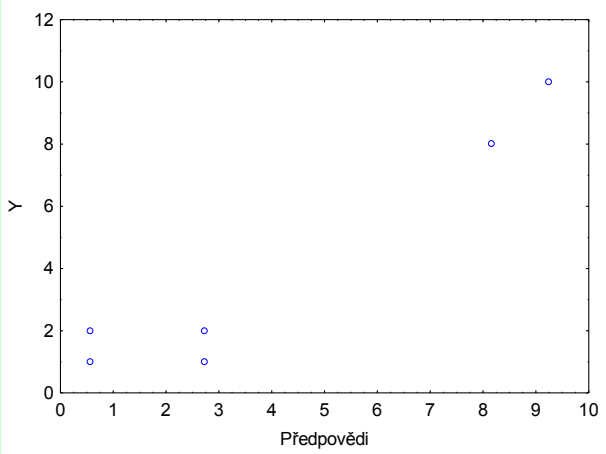
Empirická regresní funkce má tedy tvar $\hat{Y} = -1{,}09 - 0{,}38x_1 + 0{,}98x_2 + 0{,}54x_4$. Variabilita proměnné Y je z 96,5% vysvětlená zvoleným regresním modelem. Pro $\alpha = 0{,}05$ je celkový F-test nevýznamný, všechny dílčí t-testy rovněž. Podíváme-li se na beta koeficienty, vidíme, že největší vliv má proměnná X_2. Sestavíme tedy nový model $Y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i$, $i = 1, ..., 6$. Metodou nejmenších čtverců opět získáme odhady regresních parametrů.

N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,95813306 R2= ,91801897 Upravené R2= ,89752371 F(1,4)=44,792 p<,00259 Směrod. chyba odhadu : 1,2644					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			-0,520548	0,850099	-0,612338	0,573413
X2	0,958133	0,143162	1,084932	0,162108	6,692666	0,002593

Nyní má empirická regresní funkce tvar $\hat{Y} = -0{,}52 + 1{,}08x_2$, model jako celek je významný a nezávisle proměnná X_2 rovněž. Pro kontrolu kvality regrese porovnáme zjištěné a predikované hodnoty veličiny Y.

	1 student1	2 student2	3 student3	4 student4	5 student5	6 student6
Skutečnost	10.0	8.0	1.0	2.0	2.0	1.0
Predikce	9.2	8.2	2.7	2.7	0.6	0.6

Vztah mezi naměřenými a predikovanými hodnotami znázorníme pomocí dvourozměrného tečkového diagramu.



Nyní aplikujeme dopřednou metodu postupné regrese:
Statistiky – Vícerozměrná regrese – Proměnné – Závisle proměnná Y, Nezávisle proměnné X1, X2, X4 – OK – Detailní nastavení – zaškrtneme Další možnosti – OK – Metoda – zvolíme Kroková dopředná – na záložce Metoda zvolíme Zobrazit výsledky Po každém kroku – OK (V kroku 0 nejsou v regresní rovnici žádné proměnné.) Klikneme na Další – Výpočet:Výsledky regrese.

N=6	Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta) R= ,95813306 R2= ,91801897 Upravené R2= ,89752371 F(1,4)=44,792 p<,00259 Směrod. chyba odhadu : 1,2644					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			-0,520548	0,850099	-0,612338	0,573413
X2	0,958133	0,143162	1,084932	0,162108	6,692666	0,002593

V prvním kroku byla vybrána proměnná X_2. Opět klikneme na Další a dostaneme výsledky kroku 2, který je již konečný:

Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta)
R= ,97653416 R2= ,95361897 Upravené R2= ,92269829
F(2,3)=30,841 p<,00999 Směrod. chyba odhadu : 1,0981

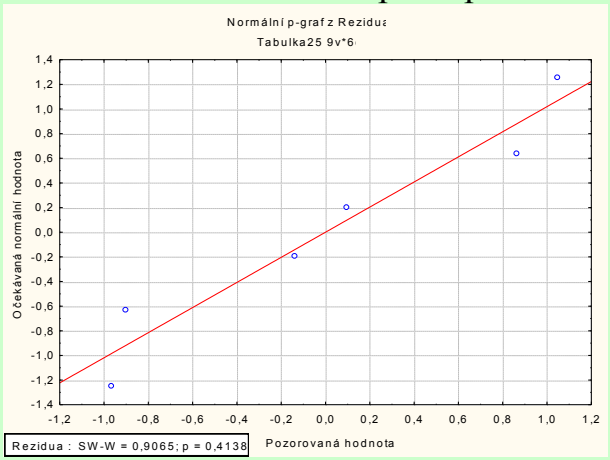
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(3)	Úroveň p
Abs.člen			-1,22615	0,872554	-1,40524	0,254603
X2	0,687789	0,217256	0,77881	0,246007	3,16580	0,050644
X4	0,329675	0,217256	0,39622	0,261109	1,51745	0,226436

Empirická regresní funkce má tvar $\hat{Y} = -1,23 + 0,78x_2 + 0,4x_4$, model jako celek je významný na hladině 0,05, avšak nezávisle proměnná X_2 a X_4 nikoliv. Přispívají však k vysvětlení variability hodnot závisle proměnné veličiny Y. Adjustovaný index determinace je 0,9227. V modelu s nezávisle proměnnou X_2 byl 0,8975 a v modelu se všemi třemi nezávisle proměnnými byl 0,9128.

V tomto výsledném modelu uložíme rezidua a predikované hodnoty:

Rezidua/předpoklady/předpovědi – Reziduální analýza – Uložit rezidua & předpovědi – OK

Pomocí S-W testu a N-P plotu prozkoumáme normalitu reziduí:



Vidíme, že rozložení reziduí je blízké normálnímu rozložení.

Zkusíme ještě zpětnou metodu postupné regrese:
Na záložce Metoda zvolíme Metoda – zvolíme Kroková zpětná. V nultém kroku jsou do modelu zařazeny všechny nezávisle proměnné:

Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta)
R= ,98240301 R2= ,96511567 Upravené R2= ,91278918
F(3,2)=18,444 p<,05187 Směrod. chyba odhadu : 1,1664

N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(2)	Úroveň p
Abs.člen			-1,08961	0,941927	-1,15679	0,366858
X1	-0,299065	0,368366	-0,38391	0,472872	-0,81187	0,502130
X2	0,864242	0,316998	0,97862	0,358949	2,72633	0,112320
X4	0,445257	0,271142	0,53513	0,325873	1,64215	0,242263

V 1. kroku je z modelu vyřazena proměnná X_1:

Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta)
R= ,97653416 R2= ,95361897 Upravené R2= ,92269829
F(2,3)=30,841 p<,00999 Směrod. chyba odhadu : 1,0981

N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(3)	Úroveň p
Abs.člen			-1,22615	0,872554	-1,40524	0,254603
X2	0,687789	0,217256	0,77881	0,246007	3,16580	0,050644
X4	0,329675	0,217256	0,39622	0,261109	1,51745	0,226436

Ve 2. kroku, který je současně poslední, je vyřazena proměnná X_4:

Výsledky regrese se závislou proměnnou : Y (ctyri testy.sta)
R= ,95813306 R2= ,91801897 Upravené R2= ,89752371
F(1,4)=44,792 p<,00259 Směrod. chyba odhadu : 1,2644

N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			-0,520548	0,850099	-0,612338	0,573413
X2	0,958133	0,143162	1,084932	0,162108	6,692666	0,002593

Metoda zpětné postupné regrese tedy jako optimální našla model regresní přímky s nezávisle proměnnou X_2.
Upozornění: Pokud bychom na záložce Metoda ručně změnili hodnoty „F na zahrnutí“ a „F na vyjmutí“, mohli bychom dostat jiné výsledky.